Early Experiences with the IBM POWER4 Scalable Parallel System at the ARL MSRC

Thomas M. Kendall and Stephen J. Schraml, U. S. Army Research Laboratory,
Aberdeen Proving Ground, MD
C. Michael McCraney, Raytheon Systems Company, Aberdeen, MD

1.0 Abstract

The IBM P-Series 690 POWER4 scalable parallel system was selected for Technology Insertion 2002 (TI-2002) at the U. S. Army Research Laboratory Major Shared Resource Center (ARL MSRC). The POWER4 architecture is revolutionary--a single chip contains dual 1.3 GHz processors with private level 1 caches and a larger shared level 2 cache. An off chip level 3 cache is shared among multiple processors. This paper discusses the effects of the shared caches, examines the absolute performance and scalability of the code, and describes some of our early experiences with the Power 4 supercomputers.

Two systems are being installed, one with 64 processors, and one with 768 processors. Each P-Series 690 consists of one cabinet or frame, and can contain up to 32 POWER4 processors. It is possible to partition each 32 CPU node into multiple logical nodes, each with its own processor(s), memory, disk and operating system. For the current study, eight processor logical partitions, each with one and two active SP switch2 Colony/Corsair PCI adapters are compared to a single 32 processor node.

Both the absolute performance and the scalability of the code will be analyzed. The CTH code is used as an application benchmark to study the effect of node (partition) size and the number of internode communication links on delivered performance. This code uses an Eulerian finite volume approach to modeling solid dynamics problems involving shock wave propagation through multiple materials with large deformations.

The Department of Defense High Performance Computing Modernization Program (HPCMP) is particularly interested in CTH performance. This code is used in millions of CPU hours each year. A large database of CTH performance already exists, allowing the results of this study to be put into perspective. CTH is written in C and FORTRAN and uses the Message Passing Interface (MPI) for parallelism. The problem size is varied to maintain nearly uniform computation to communication balance as the number of CPUs is increased.

2.0 IBM Hardware Configuration

Over the last two years, ARL has become one of IBM’s premier high performance computing sites, boasting 1384 Power3 processors in three independent environments. IBM significantly redesigned the SP nodes in the new Power4 architecture, which is offered in the P-Series 690 server. ARL proposed that IBM’s newest high performance architecture should make up the lion’s share of the TI-2002 hardware configuration. The following sections provide a brief description of the Power4 architecture, the planned ARL configurations, and an overview of the integration schedule and current status.

2.1 IBM P-series 690 Architecture

The POWER4 chip contains 2 processors—each with its own Level 1 (L1) 64KB instruction and 32 KB data caches, a shared Level 2 (L2) cache (which is just under 1.5 MB in size), and a shared directory for the off chip 0 to 32 MB (max) Level 3 (L3) cache. Four POWER4 chips (eight processors) are packaged with off chip L3 caches and system level interconnections into a multichip module (MCM). The L3 caches associated with the four chips that comprise the MCM serve as a single shared 128 MB cache.

The IBM P-Series 690 system supports up to 32 POWER4 processors in a single cabinet. It is assembled from four interconnected MCMs. These 32 procesors can be configured as a single node, or using logical partitioning, can split into multiple compute nodes--each with its own memory, disk storage, networking and operating system.

2.2 ARL Configuration and Rationale

Two independent systems were proposed and approved for the TI-2002 upgrades at ARL. A 768-processor, 768-GB system as well as a 64-processor, 96-GB system. The smaller of these two systems is targeted for the unclassified environment, initially as a development and pioneer machine, while the larger machine is targeted for the classified environment. Details of the ARL MSRC POWER4 systems are as follows:

System	1.3 GHz CPUs	Total L3 Cache	Total Memory	Logical Nodes	Total Disk
U	64	256 MB	96 GB	2-16	1.5 TB
C	768	12 GB	768 GB	24-96	7 TB

Table 1. ARL MSRC IBM P-Series 690 Configurations.

Because the P-Series architecture supports multiple virtual node configurations up to 32 processors each, the final production configuration of the 768-processor machine has yet to be determined. Upon initial delivery, each 32-processor frame will be configured as four eight processor virtual nodes. Each will have 8 GB of memory, dual 10/100 Ethernet adapters and dual Corsair/Colony switch adapters. The dual switch adapters will be configured to give individual messages striped fashion communications to increase bandwidth and reduce latency.

The system will be configured this way through capability testing and acceptance because IBM guaranteed that stated benchmarks will result. However, the integration team is also investigating non-uniform node configurations containing large memory and varying processor counts. Some ARL MSRC customers have large shared memory requirements. The 64-processor machine that is currently located in ARL's unclassified environment is ideal for these trial configurations.

Since their arrival on April 5, 2002, the two 32-processor frames have undergone several configuration changes. Initially, the frames were configured as independent 32-processor servers and a series of serial benchmark test runs were completed. Later, the two nodes received multiple reconfigurations to support the varying configurations investigated in this paper. One of the frames contains 64 GB of memory, so the ARL integration team configured this cabinet as two 16-processor nodes, each with 32 GB of memory. The second frame was configured with four, eight processor logical nodes to mirror the setup of the 768-processor machine.

In all configurations tested, each logical node always maintains dual Corsair/Colony switch adapters. Through environment variables or command line options to IBM's parallel operating environment (poe), a user can opt to use either one or two adapters per node.

Comparing operational results from each of these configurations has shown a range in performance based on consistent processor speeds, while varying node sizes and aggregate communications vs. processing performance. Table 2 lists the theoretical aggregate switch communications performance versus peak processing performance of the varying size logical nodes. Actual results from these configurations are discussed in the later sections of this document.

Node Size (# of processors)	Aggregate Processing Speed (GFLOPS)	Aggregate Switch Bandwidth (GB/s)
8	41.6	600
16	83.2	600
32	166.4	600

Table 2. Logical node performance characteristics.

2.3 ARL Integration Schedule and Status

The two P-Series 690 machines were ordered by Raytheon under the TI-2002 procurement period on 28 February 2002. By placing the order before the end of February, IBM guaranteed the systems would be delivered and operational on or before June 30, 2002. ARL elected to have the 64-processor system delivered early so that preliminary configuration tests could be performed. This was delivered on April 5, 2002.

Within seven calendar days, one of the frames was powered on and operational as a 32-processor independent server system. The Corsair/Colony adapters were received and installed three days later. While serial test runs were being completed on the first 32-processor frame, the second frame was partitioned into four logical eight processor nodes, and was operational on a single plane of the Colony switch soon thereafter.

Once the four 8-processor nodes in the second frame were operational, the original 32 processor frame was repartitioned into two 16-processor logical nodes and clustered with the other frame to form a 64 processor, 6-node system. As of this writing, all six nodes are operational on both Colony planes and share a 1.5 TB GPFS filesystem.

Because the 768-processor machine was targeted for the classified environment, ARL insisted that the machine be pre-staged at IBM's Poughkeepsie, NY facility to avoid any infant mortality or unforeseen configuration issues on site at ARL. Assembly and configuration of this system began on May 15, 2002.

At the time of this writing, the 768-processor system has been assembled, cabled, powered on, and integrated as a single, 96-node system. Stress testing is scheduled to begin the week of the HPC User's Conference, and HPCMO benchmark runs should be underway during that week.

After stress testing has been completed, the system will be disassembled, packed, and shipped to ARL. It is due to arrive by June 17, 2002. ARL and IBM anticipate an eight-day reconfiguration and setup period and an operational system as promised by June 30, 2002.

3.0 Operating Environment

For this paper, CTH was run within a single 32-processor node, across four 8-processor logical partitions, and across two 16-CPU logical partitions. The multiple partition jobs were run using unsupported Colony/Corsair adapters and unsupported software. The results from these runs are not expected to be fully representative of the released system. It is expected that the released firmware and software will improve upon the performance of the current system.

The nodes were all configured with AIX 5, PSSP 3.4, FORTRAN compiler version 7.1.1, and C/C++ compiler version 5.0.2.

4.0 Applications

CTH (McGlaun and Thompson 1990) is an Eulerian finite volume code for the modeling of shock wave propagation through multiple materials in one, two, or three dimensions, producing large deformations. The code employs a two-step scheme, where a Lagrangian step is followed by a remap step. The conservation equations are replaced by explicit finite volume equations that are solved in the Lagrangian step. Operator splitting techniques are employed to replace multidimensional equations with a set of one-dimensional equations in the remap step.

CTH is implemented using a single program multiple data (SPMD) methodology, which is one of several that have been shown to be effective on scalable computing architectures. Under the SPMD paradigm, the same executable code runs on each CPU but operates on a different set of data. Algorithms such as the ones used in CTH are well suited for SPMD methodology because the problem domain can be divided into fixed subdomains that each reside on a CPU.

Ghost cells are used to adopt CTH to the SPMD paradigm. Explicit message passing is used to exchange data between neighboring subdomains by way of these cells. A thorough description of the distributed finite volume approach used in CTH is available in Kimsey et. Al. (1998).

5.0 Problem Description

CTH has been used to model a long rod projectile impacting an oblique steel plate. This problem was chosen because a rich set of well-characterized, experimental data exists--as well as a set of serial computational results by Hertel (1992).

Fugelso and Taylor (1978) conducted a series of ballistic experiments to evaluate the effects of combined obliquity and yaw on high density long rod projectiles. Depleted uranium (DU) alloy long rod projectiles with little of no yaw were launched into an oblique rolled homogeneous armor (RHA) plate that had been accelerated by an explosive charge, resulting in a yawed impact in the plate frame of reference.

The DU projectiles were right circular cylinders with a hemispherical nose and impact velocities ranged from 850 to 1650 m/s. The length and diameter of the projectile in Shot 58 of the test series were 7.67 cm and 0.767 cm, respectively, for a length to diameter ratio of 10. The RHA plate thickness was 6.4 mm. The striking velocity was 1289 m/s. Figure 1 shows a schematic of the test initial conditions.

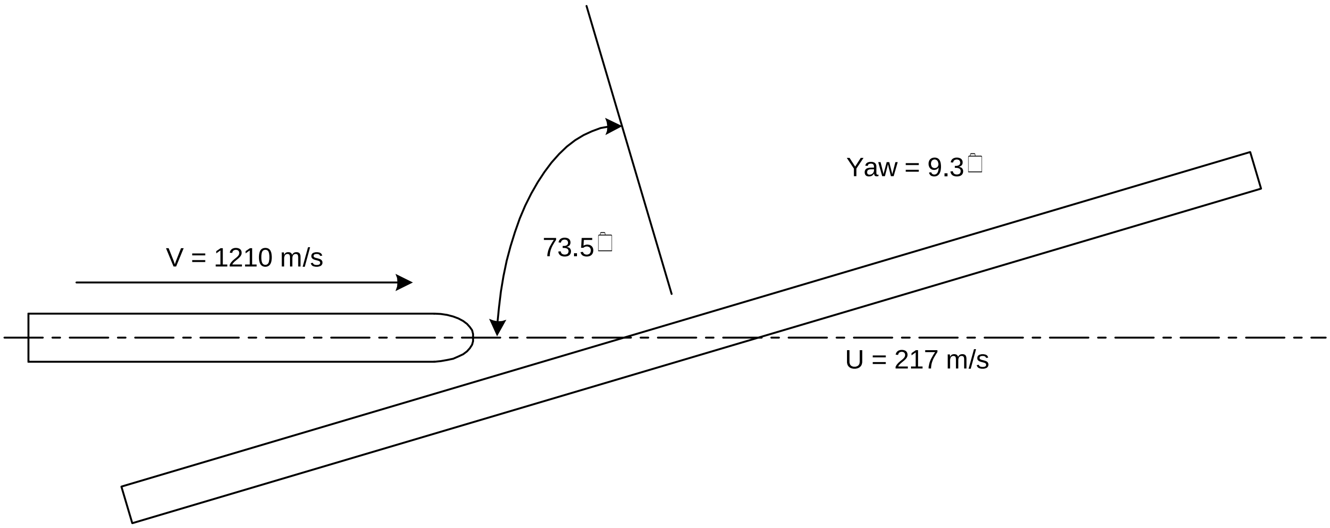

Figure 1. Initial Conditions for the combined yaw and obliquity impact simulation.

The workload per processor was kept as uniform as possible constant to maintain a uniform computation to communication ratio. During the timed run, minimal disk access was performed, so only computation and communications performance are measured. All computations simulate 40 μs of time.

The grid was refined by uniformly decreasing the characteristic length by a factor of $2^{-1/3}$ as the number of processors are doubled. This doubles the total number of grid points, maintaining the computation to communication ratio at a uniform level.

The characteristics of the grids used in this study are shown in Table 2. The columns NI, NJ, and NK refer to the number of Eulerian cells in the *x*, *y*, and *z* directions, respectively, excluding ghost cells. For the 128-processor simulation, almost 50 million cells are used.

Number of Processors	NI	NJ	NK	Zone Length (mm)	Total Zones
1	215	30	60	1.000	387,000
2	271	38	75	0.794	772,350
4	341	48	95	0.630	1,554,960
8	430	60	120	0.500	3,096,000
16	541	76	151	0.397	6,208,516
32	683	95	191	0.315	12,393,035
64	860	120	250	0.250	24,768,000
128	1083	151	302	0.198	49,386,966

Table 3. Description of the computational grids used in the scalability studies.

6.0 Scalability Results

Three configurations of P-Series 690 systems were evaluated:

- A single 32-processor node
- A set of 8-processor logically partitioned nodes with a single Corsair adapter were used with data ranging from 1 to 128 processors by powers of 2

- The second data set was repeated using two active Corsair adapters per 8-processor logical partition

The data from the scalability study on the 32-processor node are shown in Figure 2. The circles represent the measured grind times and the solid line represents the ideal scalability extrapolated from the single processor run. The grind time is defined as the mean time to perform differencing on one Eulerian cell in one computational cycle.

Given the linear relationship between the measure grind time and the number of processors used, the scalability can be described by the equation $g_n = g_1 / n^m$, where g_n is the predicted grind time for *n* processors, g_1 is the measured grind time from the single processor simulation, and *m* is the parallel efficiency.

Regression analysis is used to obtain the value of *m* for a given set of measured data. A value of *m* equal to 1 represents ideal scalability. For the 32-processor node, the value of *m* was determined to be 0.8374.

The dashed line in Figure 2 represents the computed scalability trend. The difference between the two lines shows the amount the measured scalability diverges from ideal scalability.

The two- and four-processor simulations are likely benefiting substantially from the shared cache architecture. Once eight of the 32 processors are used, the performance degrades somewhat as the L3 cache is now split among at least two processors.

The 32-processor calculation shows further divergence the ideal. At this point, all L2 and L3 caches are being shared by all of their attached processors, resulting in the minimum bytes available per CPU of all of the simulations performed. Even with the effects of the shared caches benefiting the smaller processor count simulations, it is encouraging that the computed slope from the 32-processor P Series 690 is nearly equal to a single 16-processor IBM Power3 node (*m*=0.8477, Schraml et. al. 2001).

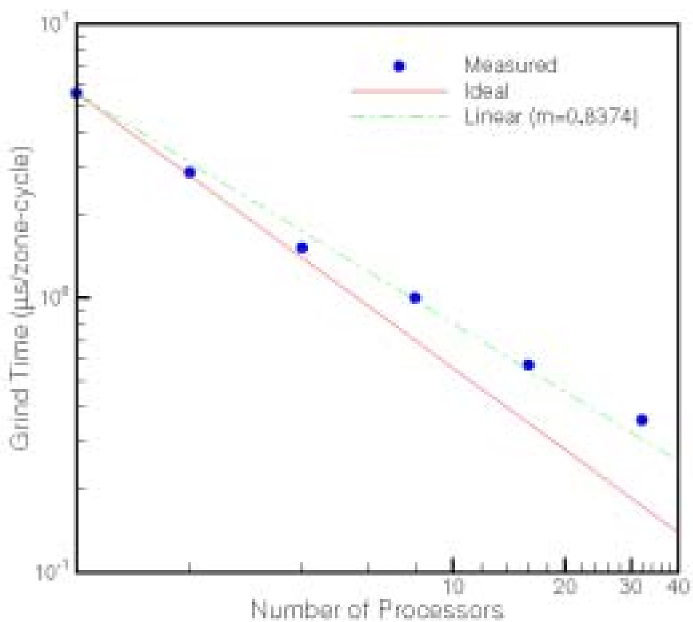

Figure 2. Scalability of CTH within a single 32 processor IBM Series 690.

The results for the second configuration are shown in Figure 3. This system was configured with eight processors and a single active Corsair adapter. The slope of the curve for this case was *m*=0.7259, which is substantially less than the slope of the curve for the runs on the single 32 processor node.

The two-processor through sixteen- processor cases are run using two nodes and the others are run on nodes filled with eight MPI tasks each. Again, the smaller processor count simulations are likely benefiting from the higher amounts of cache available per MPI task. The overall scalability is far less ideal than the single node case.

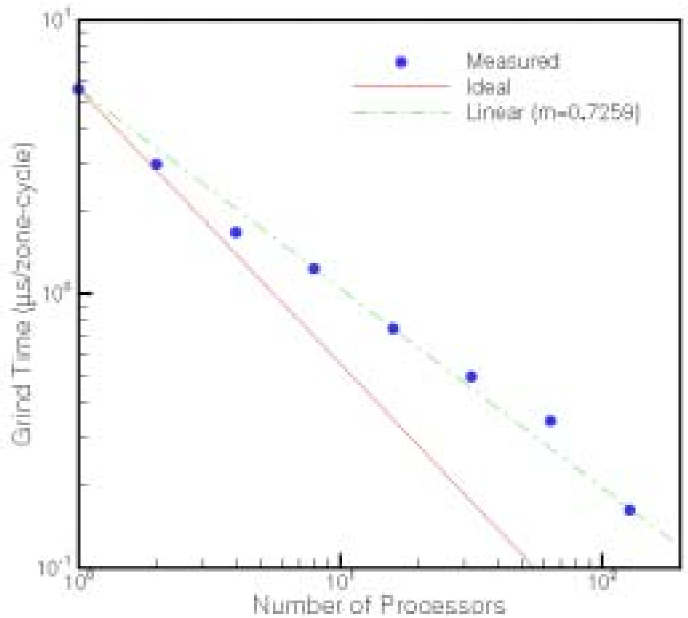

Figure 3. Scalability of CTH on the IBM P-Series 690 using one Corsair adapter per 8-processor logical partition.

The third tested configuration also used eight-processor logical nodes. Each node had dual Corsair adapters, improving the balance between the system's communications and computational capabilities. The results from this case are shown in Figure 4. The slope improves by 6% from the single corsair case to 0.7715. This increase in scalability leads to a forecasted improvement of 35% for a 768-processor job. The effects of the shared caches are evident once again.

While the addition of the second Corsair adapter provided a significant improvement over the single adapter study, the scalability shows more divergence from ideal than other systems recently tested. The IBM Power3 RS/6000 SP with a single Colony adapter produced a slope of 0.844, while the SGI Origin3800 produced a value of 0.878 (Schraml et. al., 2001). However, the software support for dual Corsair adapters was not yet released at the time of the simulations. The progress towards improving the overall scalability will be closely monitored.

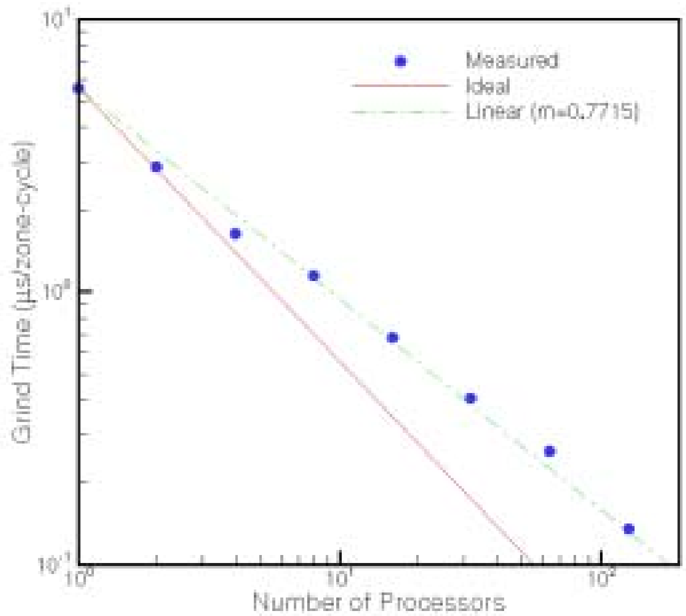

Figure 4. Scalability of CTH on the IBM P-Series 690 using one corsair adapter per 8-processor logical partition.

7.0 Conclusions

The scalability of the CTH hydrodynamics code on a new microprocessor and system architecture were studied using a nearly constant number of zones per processor. The results indicate that the serial performance of the IBM P-Series 690 is very impressive and delivers slightly greater than a three-fold increase over its predecessor, the IBM RS/6000 SP with the 375 MHz POWER3 microprocessor.

The scalability within a single node is also very impressive. While there is room for improvement in the internodal communications performance, the introduction of support for dual switch adapters per node shows significant improvement over the single switch adapter configuration. The IBM P-Series 690 system is certainly an excellent platform for large-scale continuum simulations.

8.0 Acknowledgements

The authors wish to acknowledge Michael Cook of IBM for his assistance in performing the simulations. We also wish to thank George Kraft of Raytheon Systems Company and Bill Wells of the Open Technology Group for their contributions towards configuring the system.

9.0 References

http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.pdf

Hertel, E. S. "A Comparison of the CTH Hydrodynamics Code With Experimental Data." SAND92-1879, Sandia National Laboratories, Albuquerque, NM, 1992.

Fugelso, E. and Taylor, J. W. “Evaluation of Combined Obliquity and Yaw for U-0.75% Ti Penetrators.” LA-7402-MS, Los Alamos National Laboratories, Los Alamos, NM, 1978.

Schraml, S. J., Kimsey, K. D., and Kendall T. M. “Scalable Simulations of Penetration Mechanics on the SGI Origin3800 and the IBM SP POWER3 Computer Systems, ARL-TR-2537, U.S. Army Research Laboratory, 2001.